Cross-Validation

Problem

Split the dataset $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N]$ of $\boldsymbol{N}$ objects, each of which belongs to one of the $\boldsymbol{M}$ classes ($\forall \boldsymbol{i}$: $1 \le \boldsymbol{a}_i \le \boldsymbol{M}$), into $\boldsymbol{K}$ disjoint parts. Each object must fall into exactly one part. Splitting must satisfy several constraints.

Sizes balance constraint

Let cnt($\boldsymbol{s}$) is the number of objects belonging to the $\boldsymbol{s}$ part, then

$$\forall \boldsymbol{s},\boldsymbol{t}: |\text{cnt}(\boldsymbol{s}) - \text{cnt}(\boldsymbol{t})| \le 1$$

Classes distribution balance constraint

Let cnt($\boldsymbol{s},\boldsymbol{c}$) is the number of objects with class $\boldsymbol{c}$ belonging to the $\boldsymbol{s}$ part, then

$$\forall \boldsymbol{s},\boldsymbol{t},\boldsymbol{c}: |\text{cnt}(\boldsymbol{s},\boldsymbol{c}) - \text{cnt}(\boldsymbol{t},\boldsymbol{c})| \le 1$$

Randomness of splitting constraint

All possible splitting should be equiprobable. To be sure in that, for the same input run splitting several times and check that:

1. For each object probability of being in any part is equally distributed.
2. For each pair of objects probability of being together in the same part is equally distributed.

Example

Let $\boldsymbol{N}$ = 10, $\boldsymbol{M}$ = 4, $\boldsymbol{K}$ = 3, and

$\boldsymbol{A} = [\boldsymbol{a}_0 = 1, \boldsymbol{a}_1 = 1, \boldsymbol{a}_2 = 1, \boldsymbol{a}_3 = 1, \boldsymbol{a}_4 = 2, \boldsymbol{a}_5 = 2, \boldsymbol{a}_6 = 2, \boldsymbol{a}_7 = 3, \boldsymbol{a}_8 = 3, \boldsymbol{a}_9 = 4]$

Then:

- The splitting $\boldsymbol{S}_1 = [\boldsymbol{a}_0,\boldsymbol{a}_1,\boldsymbol{a}_2]$, $\boldsymbol{S}_2 = [\boldsymbol{a}_3,\boldsymbol{a}_4,\boldsymbol{a}_5]$, $\boldsymbol{S}_3 = [\boldsymbol{a}_6,\boldsymbol{a}_7,\boldsymbol{a}_8,\boldsymbol{a}_9]$ satisfy the first constraint but it isn't satisfy the second.
- The splitting $\boldsymbol{S}_1 = [\boldsymbol{a}_0,\boldsymbol{a}_1,\boldsymbol{a}_4,\boldsymbol{a}_7,\boldsymbol{a}_9]$, $\boldsymbol{S}_2 = [\boldsymbol{a}_2,\boldsymbol{a}_5,\boldsymbol{a}_8]$, $\boldsymbol{S}_3 = [\boldsymbol{a}_3,\boldsymbol{a}_6]$ satisfy the second constraint but it isn't satisfy the first.
- The splitting $\boldsymbol{S}_1 = [\boldsymbol{a}_0,\boldsymbol{a}_1,\boldsymbol{a}_4,\boldsymbol{a}_7]$, $\boldsymbol{S}_2 = [\boldsymbol{a}_2,\boldsymbol{a}_5,\boldsymbol{a}_8]$, $\boldsymbol{S}_3 = [\boldsymbol{a}_3,\boldsymbol{a}_6,\boldsymbol{a}_9]$ satisfy the both constraints.